DOCUMENT RESUME

ED 042 572 RE 002 835

AUTHOR Dewey, Godfrey
TITLE Relative Frequency of Occurrence as a Factor in the Phonemic and Graphemic Problems of Written English.
PUB DATE 7 May 70
NOTE 11p.; Paper presented at the conference of the International Reading Association, Anaheim, Cal., May 6-9, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.65
DESCRIPTORS English, Graphemes, *Linguistics, *Orthographic Symbols, Phonemes, *Phonemic Alphabets, *Phonemics, *Reading, Spelling

ABSTRACT

Two criteria for a phonemic notation--assignment of symbols to sounds and the influence of purpose--were discussed. Also presented were three purposes of a phonemic notation: (1) as an initial teaching medium, (2) for an intermediate stage or stages of phonemic spelling reform of English, and (3) for an ultimate phonemics spelling reform. Data of relative frequency of phonemes and/or graphemes were viewed as having importance in (1) devising phonemic codes and in formulating rules, (2) assessing compatibility with traditional orthography, (3) estimating the possible savings in the writing and printing of superfluous letters, and (4) standardizing a type of notation. Examples were given which illustrate the use of data on relative frequency in making decisions related to standardizing the type of notation. Tables and references are included. (DH)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

POOR ORIGINAL COPY - BEST AVAILABLE AT TIME FILMED

4/4/70 edition

ED042572

# Relative frequency of occurrence as a factor in the phonemic and graphemic problems of written English

Godfrey Dewey

(IRA-SSA meeting, Anaheim, Thursday, May 7, 1970)

Strictly speaking, the announced title of this paper should have included one more word, referring specifically to problems of written English. The English language which we speak is no more affected by whether it is recorded graphemically in shorthand or longhand, typing or print, than by whether it is recorded acoustically on a cylinder, a disk, or a tape. The very first sentence of the classic "Principls of '76" (I retain the original spelling) promulgated by the American Philological Association in 1876 was:

> The true and sole office of alfabetic writing is faithfully and intelligibly to represent spoken speech.[1]

and it is with the written representation that we are here chiefly concerned.

Criteria for a phonemic notation of whatever type may be grouped in four main categories: sounds, symbols, assignment of symbols to sounds, and the influence of purpose.[2] For each of these categories, statistics on relative frequency of phonemes and/or graphemes are significant in varying degree. In the limited time available for this paper, the first two will have to taken for granted; assuming substantially the phonemic basis of i.t.a. and WES (World English Spelling), and the graphemic basis of WES, either of which would be a major topic in itself, and confining our examination to the third, assignment of symbols to sounds, as modified by the fourth, the influence of purpose.

Data on relative frequencies here cited are, unless otherwise specified, taken from my studies of phonemes[3] and graphemes[4], both based on exhaustive analysis of the same 100,000 words of well-diversified connected matter, on a 41-phoneme basis (counting schwa), virtually equivalent to the phonemic basis

RE002 835

POOR ORIGINAL COPY - BEST AVAILABLE AT TIME FILMED



of i.t.a. Complete data on occurrences and items are stated usually in the form x/y, where x equals the total of occurrences on the printed page, and y equals the number of items (different words or syllables) involved--per 100,000 running words always understood. In general, data on occurrences are more significant for reading, data on items more significant for writing, i.e., spelling.

Statistics, however carefully compiled, are chiefly valuable as an aid to common sense, not as a substitute for it. In particular, decisions should never be based on the most frequent spellings of sounds without taking into account the most frequent pronunciations of spellings. These are not just inverted statements of the same fact. Thus the predominant spellings of the name-sounds of A, E, U are the letters a, e, u, but the predominant pronunciations of the letters a, e, u are as in bat, bet, but respectively. Similarly, the commonest spelling of the phoneme /z/ is the letter s, but the commonest pronunciation of the letter s is /s/.

Data on relative frequency of phonemes and/or graphemes can be invaluable both in devising phonemic codes and in formulating rules and/or exceptions for their practical application. I say codes, rather than a code, because as of today no one phonemic code for English can conceivably be "best" for all purposes. At the phonemic level, setting aside the precise phonetic notations which are the legitimate and valuable tools of the linguistic scholar, but a perplexing mystery to the untrained ear, there are at least three somewhat different purposes to be served by a phonemic notation: 1) As an i.t.m. (initial teaching medium), the purpose of most immediate interest to us; 2) For an intermediate stage or stages of phonemic spelling reform of English; 3) For an ultimate phonemic spelling reform.

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED



Much of the importance of data on relative frequency derives from the problem of compatibility with T.O. (traditional orthography). For an i.t.m., the importance of compatibility in facilitating the all-important transition to T.O. is sufficiently obvious. For an intermediate stage of spelling reform, to be used, as Shaw put it, "side by side with the present lettering until the better ousts the worse,"[5] the necessity for an essentially "self-reading" degree of compatibility, for one who has never examined the code, is no less obvious. Even for an ultimate spelling reform, which in the English-speaking countries could hardly be imposed by decree, as Kemal Ataturk imposed the Roman alphabet on Turkish, compatibility would surely minimize resistance to the transition.

Yet another point at which data on relative frequency make a significant contribution is in estimating the possible savings in the writing and printing of superfluous letters--the aspect on which Shaw again and again laid extravagant emphasis.[6] For a well-designed phonemic alphabet of the supplementing type (one sign, one sound, adding necessary new letters to the present Roman alphabet), this saving can run just about 1 letter in 6, or $170,000,000 out of each $1,000.000,000 of writing and printing costs. For the more immediately practicable standardizing (no-new-letter) type, the difference from T.O. will be only 1 or 2% either way, since the necessary new digraphs, chiefly for the long vowels and diphthongs, just about offset the saving of silent or otherwise superfluous letters.

For the purpose of most immediate interest to us, initial teaching media, WES will serve to supply examples of the application of relative frequency data to the standardizing type of notation. The supplementing type, of which

i.t.a. is the prime exemplar, involves too many subjective judgments as to the degree of compatibility of characters not now in the Roman alphabet to be dealt with statistically in a paper of this length. As oral presentation of comparative figures on phonemes and graphemes is not easy to follow, three exhibits have been provided: "World English Spelling (WES) for better reading"; the SSA (Simpler Spelling Association) Phonemic Alphabet, which most nearly parallels the phonemic basis of WES; and selected pre-publication figures on relative frequency of spellings,[7] to which I have added, for ease of oral presentation, figures for percentages of occurrences, rounded off to the nearest 1%.

If compatibility is to be regarded as the predominant criterion, the Roman alphabet letters for about half of the consonant phonemes and most of the short vowel phonemes call for no comment. Because of the awkwardness of oral presentation, the examples discussed will be confined to a few of the most difficult or controversial decisions, both consonant and vowel: for consonants, the th problem, and the treatment of c and g and s; for vowels, the "u" group of phonemes, as in but, full, fool; and an examination of the three principal differences between WES as a spelling reform notation and as an i.t.m.

If only items are considered, the all-too-common practice in the past, it appears that the th grapheme is pronounced unvoiced, as in thin, 65% of the time, voiced, as in then, only 35%. This leads naturally to assigning the familiar th grapheme to the unvoiced phoneme, with the logically cognate but uncouth symbol dh for the voiced phoneme. If, however, occurrences, the more appropriate criterion for reading, be considered, it appears that 90% of all occurrences are pronounced with the voiced sound, so that assignment of the th grapheme to other than the voiced phoneme is unthinkable. In that case,

however, there remains no satisfactory digraph for the unvoiced phoneme. The cumbersome but intelligible thh grapheme adopted in WES may be justified to a degree by relative frequency data on two grounds: 1) the phoneme is one of the four least frequent in English, only 0.37%; 2) for native English-speaking users, the distinction is virtually unnecessary. In the entire 17,000 different words of the recent Hanna study,[8] derived chiefly from the 4.5 million running words which formed the basis of the Thorndike-Lorge list,[9] there are only 6 pairs of words (ether, either; thigh, thy; loath, loathe; mouth, mouth; sheath, sheathe; wreath, wreathe) distinguished phonemically only by surd or sonant pronunciation of th; and of these, only one word (either), no pair, occurs in my list of commonest words, which includes all those found oftener than once in 10,000 running words.

Use (or non-use) of the grapheme c is bound up with the phonemes /k/ and /s/. /k/ is spelled c in 64% of all occurrences, k in 18%, and 9 other ways totaling 18%. Conversely, however, k is pronounced /k/ in all occurrences, whereas c is pronounced /k/ in only 72% of occurrences, /s/ in 28%. Thus, explicitness, as well as the more distinctive form of the letter, obviously calls for representing /k/ by k.

A parallel example is the phoneme /j/, which is spelled g in 60% of all occurrences, j in 26%, and 8 other ways totaling 14%. Again, however, j is pronounced /j/ in all occurrences, whereas g is pronounced /g/ in 73% of all occurrences, but /j/ in only about 27%, with 3 other ways totaling less than 0.5%. Quite obviously, therefore, explicitness calls for representing /j/ by j, and restricting g to /g/--except, of course, for the digraph ng which, like any digraph, is regarded as a unitary symbol.

One more example of the importance of considering pronunciations as well as spellings in order to maintain the "self-reading" quality which is one factor in compatibility. The phoneme /s/ is spelled s in 75% of all occurrences, c in only 14%, and 7 other ways totaling 11%. The letter s, however, is pronounced /s/ in only 54% of all occurrences, /z/ in 45%, and 2 other ways totaling 1%. Conversely, the phoneme /z/ is spelled s in 97% of all occurrences; the letter z is pronounced /z/ in 96% of all occurrences. This preponderance conclusively calls for representing/z/ by z, leaving s as the explicit representation for /s/.

Assignment of graphemes for the three vowel phonemes spelled oo in food, good, and flood, is a particularly good example of the help which relative frequency data can render. It will be taken for granted that the best available graphemes are oo, uu (which does not occur in T.O. but is used in the British New Spelling), and u; for discussion of the reasons for eliminating dual use of w as a vowel, or employing some digraph including w, or whatever, would range too far outside the scope of this paper.

The vowel phoneme in food is spelled o in 60% of all occurrences (which however includes the preposition to (2924/1, 48%) most commonly pronounced with schwa, ou in 19%, oo in 7%, and 15 other ways totaling 11%. Conversely, oo is pronounced /ɯ/ in 50% of all occurrences, /u/ in 45%, /o/ in 3%, and /ʊ/ in 2%.

The vowel phoneme in good is spelled u in 24% of all occurrences, ou in 21%, oo in 15%, o in 15%, and 7 other ways totaling 25%. For comparison, ou is pronounced /ɑɯ/ in 38% of all occurrences, /ɯ/ in 30%, /u/ in 14%, /ʊ/ in 14%, /o/ in 3%, and /ə/ in 1%.

The vowel phoneme of flood is spelled u in 60% of all occurrences, o in 14%, ou in 8%, oo in less than 0.5%, and 6 other ways totaling 18%. Conversely, u is pronounced /ʊ / in 64% of all occurrences, /u / in 10%, /ɯ / in 8%, and 5 other ways totaling 18%.

Correlating the foregoing figures--

For assignment of the oo grapheme, the fact that 50% of its occurrences are pronounced /ɯ /, as against 45% pronounced /u /, is hardly conclusive. When however it is noted that the commonest spelling of /ɯ / is o, and the commonest spelling of /u/ is u, the preponderance of the evidence clearly favors assignment of oo to /ɯ /. Since the predominant spelling of /ʊ / is u, and the predominant pronunciation of u is /ʊ /, the traditional assignment of u to "short u" is fully confirmed. This leaves uu as the inevitable and not inappropriate choice for the phoneme /u/, most commonly referred to, or keyed in diacritic notations, as "short oo."

## Concessions from one sound, one symbol writing

In principle, the chief distinction between a spelling reform notation and an i.t.m. lies in striking the balance betwee maximum simplicity (i.e., regularity) and maximum compatibility with T.O.[10] In practice, relative frequency data support three major concessions from one symbol for one sound writing (not, be it noted, from one sound for one symbol) introduced by i.t.a. and paralleled by WES.

1) Doubled consonants for a single phoneme, where T.O. has doubled consonants. Of the 21 consonant letters of the Roman alphabet (counting the semivowels, w, y, h), 6 (h, k, q, w, x, y) apparently are not doubled in T.O., and 2 more (jj, vv) did not occur in the 100,000 running words which I examined.

The remaining 13, plus ck (in effect a doubled consonant)--bb, cc, ck, dd, ff, gg, ll, mm, nn, pp, rr, ss, tt, zz--occur 7070/1656 times, of which 99% represent the same phoneme assigned to the corresponding single consonant. In consequence, retention of these occurrences improves the compatibility of some 6,900 running words in 100,000, and preserves the exact T.O. forms of some 2,000; at the same time that it introduces a simple but significant step toward the eventual transition to T.O.

2) Writing c for /k/, where T.O. has c for /k/; including cc and ck. The figures for /k/ and c, showing /k/ spelled 64% by c, and c pronounced 72% as /k/, have already been cited. This concession improves the compatibility of some 6,500 words, and preserves the exact T.O. forms of some 1,200; and again builds another simple bridge toward the impending transition to T.O.

3) Writing y for the high front unstressed vowel (between /i/ and /ɛ/) which Sir James Pitman has aptly named schwi, where T.O. writes y for that sound at the end of a word or root. The accompanying exhibits, showing /i/ spelled y in 14% of occurrences, and y pronounced /i/ (in most cases, schwi), in 61% of all occurrences, speak for themselves. This concession improves the compatibility of some 4,666 words, and preserves the exact T.O. forms of some 800; again, building toward the transition to T.O.

To take full advantage of data on relative frequency of phonemes and graphemes is a far more intricate problem than these relatively simple and straightforward examples might seem to indicate. For example, nothing has been said on the problem of selecting the most suitable digraphs, and only one example has been given of their assignment. Enough has been said, however, I hope, to indicate the importance of the relative frequency aspect in practical linguistics.

## References

1. March, Francis. The spelling reform. U. S. Bureau of Education, Circular of information No. 8, 1893. Washington: Government Printing Office, 1893, p. 16.

2. Dewey, Godfrey. English spelling: Roadblock to reading. New York: Teachers College Press (In preparation), Appendix C.

3. Dewey, Godfrey. Relativ frequency of English speech sounds. Cambridge: Harvard University Press, 1923, 1950.

4. Dewey, Godfrey. Relative frequency of English spellings. New York: Teachers College Press, 1970.

5. Shaw, George Bernard. Preface to R.A. Wilson, The miraculous birth of language. London: J.M. Dent & Sons, Ltd., 1942, p. xxxi.

6. Tauber, Abraham. George Bernard Shaw on language. New York: Philosophical Library, 1963, especially p. 65-136.

7. See #4, Tables 5 and 6

8. Hanna, Paul R., et al. Phoneme-grapheme correspondences as cues to spelling improvement. Washington: U.S. Government Printing Office, 1966. (Doc. OE-32008)

9. Thorndike, Edward L., and Irving Lorge. The teacher's wordbook of 30,000 words. New York: Teachers College, Columbia University, 1944.

10. See #2.

Relative frequency of occurrence
as a factor in the phonemic and graphemic problems of written English

by Godfrey Dewey

Cosponsored meeting, I.R.A. and S.S.A., Thursday, May 7, 1970, 4:00-5:00 p.m.

Pre-publication data from Relative frequency of English spellings, by Godfrey Dewey

## C o n s o n a n t s

| Spellings of phonemes | | | | | Pronunciations of graphemes | | | | |
|---|---|---|---|---|---|---|---|---|---|
| /ŧ/ | th | 12,757*/ | 114* | 90% / 35% | th | /ŧ/ | 12,757*/ | 114* | 90% / 35% |
| /ŧ/ | th | 1,392 / | 212 | 10%-/ 65% | | /ŧ/ | 1.392 / | 212 | 10% / 65% |
| | h | 4 | 1 | 0 0 | | | 14,149*/ | 326* | |
| | | 14,153*/ | 327* | | | | | | |
| /k/ | c | 6,403 / | 1775 | 64% | k | /k/ | 6,403 / | 1775 | 100% |
| | k | 1,854 / | 343 | 18% | | | | | |
| 9 others | | 1,753 / | 562 | 18% | c | /c/ | 6,403 / | 1775 | 72% |
| | | 10,010 / | 2680 | | | /s/ | 2,477 / | 622 | 28%- |
| | | | | | | /ʃ/ | 17 / | 11 | 0 |
| | | | | | | | 8,897 / | 2408 | |
| /j/ | g | 948 / | 306 | 60% | j | /j/ | 414 / | 111 | 100% |
| | j | 414 / | 111 | 26% | | | | | |
| 8 others | | 220 / | 75 | 14% | g | /g/ | 2,616 / | 560 | 73% |
| | | 1,582 / | 492 | | | /j/ | 948 / | 306 | 27%- |
| | | | | | | /ʒ/ | 6 / | 5 | 0 |
| | | | | | | | 3,570 / | 871 | |
| /s/ | s | 12,822 / | 2974 | 75% | s | /s/ | 12,822 / | 2974 | 54% |
| | c | 2,477 / | 622 | 14% | | /z/ | 10,695 / | 1902 | 45% |
| 7 others | | 1,782 / | 566 | 11% | 2 others | | 136 / | 30 | 1% |
| | | 17,081 / | 4162 | | | | 23,653 / | 4906 | |
| /z/ | s | 10,695 / | 1902 | 97% | z | /z/ | 247 / | 107 | 96% |
| | z | 247 / | 107 | 2% | 2 others | | 9 / | 6 | 4% |
| 5 others | | 147 / | 54 | 1% | | | 256 / | 113 | |
| | | 11,089 / | 2063 | | | | | | |

*Includes the 7,310 / 1

## Vowels

| Spellings of phonemes | | | | | Pronunciations of graphemes | | | | |
|---|---|---|---|---|---|---|---|---|---|
| /ɯ/ | o | 3,645*/ | 26* | 60% | oo | /ɯ/ | 430 / | 88 | 50% |
| | ou | 1,127 / | 36 | 19% | | /u/ | 388 / | 54 | 45% |
| | oo | 430 / | 88 | 7% | | /o/ | 27 / | 6 | 3% |
| | u | 161 / | 48 | 3% | | /ʊ/ | 17 / | 7 | 2% |
| 15 others | | 688 / | 124 | 11% | | | 862 / | 155 | |
| | | 6,051*/ | 322* | | | | | | |
| /u/ | u | 604 / | 171 | 24% | ou | /ɑ/ | 1,422 / | 150 | 38% |
| | ou | 546 / | 8 | 21% | | /ɯ/ | 1,127 / | 36 | 30% |
| | oo | 388 / | 54 | 15% | | /u/ | 546 / | 8 | 14% |
| | o | 368 / | 14 | 15% | | /ʊ/ | 527 / | 157 | 14% |
| 7 others | | 671 / | 219 | 25% | | /o/ | 117 / | 21 | 3% |
| | | 2,577 / | 466 | | | /ə/ | 22 / | 11 | 1% |
| | | | | | | | 3,761 / | 383 | |
| /u/ | u | 3,768 / | 797 | 60% | u | /u/ | 3,768 / | 797 | 64% |
| | o | 857 / | 104 | 14% | | /ʊ/ | 604 / | 171 | 10% |
| | ou | 527 / | 157 | 8% | | /ɯ/ | 498 / | 186 | 8% |
| | oo | 17 / | 7 | 0 | 5 others | | 1,039 / | 279 | 18% |
| 6 others | | 1,104 / | 53 | 18%- | | | 5,909 / | 1433 | |
| | | 6,273 / | 1118 | | | | | | |

## Concession

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| /i/ | i | 20,276 / | 3807 | 69% | i | /i/ | 20,276 / | 3807 | 89% |
| | y | 4,100 / | 885 | 14% | | /ɑ/ | 2,107 / | 302 | 9% |
| | e | 2,833 / | 803 | 10% | 3 others | | 491 / | 101 | 2% |
| 17 others | | 2,074 / | 467 | 7% | | | 22,874 / | 4210 | |
| | | 29,283 / | 5962* | | | | | | |
| /y/ | y | 1,507 / | 40 | 67% | y | /i/ | 4,100 / | 885 | 61% |
| | l | 145 / | 36 | 6% | | /y/ | 1,507 / | 40 | 22% |
| 4 others | | 608 / | 174 | 27% | | /ɑ/ | 1,154 / | 73 | 17% |
| | | 2,260 / | 250 | | | /ə/ | 1 / | 1 | 0 |
| | | | | | | | 6,762 / | 999 | |

*Includes the preposition to, 2,924/1, 48%; most commonly pronounced with /ə/